Learning From Compressed Observations

Maxim Raginsky
Beckman Institute and University of Illinois
405 N Mathews Ave, Urbana, IL 61801
maxim@uiuc.edu

***Abstract*— The problem of statistical learning is to construct a predictor of a random variable Y as a function of a related random variable X on the basis of an i.i.d. training sample from the joint distribution of (X,Y). Allowable predictors are drawn from some specified class, and the goal is to approach asymptotically the performance (expected loss) of the best predictor in the class. We consider the setting in which one has perfect observation of the X-part of the sample, while the Y-part has to be communicated at some finite bit rate. The encoding of the Y-values is allowed to depend on the X-values. Under suitable regularity conditions on the admissible predictors, the underlying family of probability distributions and the loss function, we give an information-theoretic characterization of achievable predictor performance in terms of conditional distortion-rate functions. The ideas are illustrated on the example of nonparametric regression in Gaussian noise.**

I. Introduction and problem statement

Let X and Y be jointly distributed random variables, where X takes values in an *input space* $\mathcal{X}$ and Y takes values in an *output space* $\mathcal{Y}$. The problem of statistical learning is about constructing an accurate predictor of Y as a function of X on the basis of some number of independent copies of (X,Y), often with very little or no prior knowledge of the underlying distribution. A very general decision-theoretic framework for learning was proposed by Haussler [1]. In a slightly simplified form it goes as follows. Let $\mathcal{P}$ be a family of probability distributions on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Each member P of $\mathcal{P}$ represents a possible relationship between X and Y. Also given are a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ and a set $\mathcal{F}$ of functions (*hypotheses*) from $\mathcal{X}$ into $\mathcal{Y}$. For any $f \in \mathcal{F}$ and any $P \in \mathcal{P}$ we have the *expected loss* (or *risk*)

$$L(f,P) = \mathbb{E}\,\ell(f(X),Y) \equiv \int_{\mathcal{Z}} \ell(f(x),y)dP(x,y),$$

which expresses quantitatively the average performance of f as a predictor of Y from X when $(X,Y) \sim P$. Let us define the minimum expected loss

$$L^*(\mathcal{F},P) \stackrel{\triangle}{=} \inf_{f\in\mathcal{F}} L(f,P)$$

and assume that the infimum is achieved by some $f^* \in \mathcal{F}$. Then f^* is the best predictor of Y from X in the hypothesis class $\mathcal{F}$ when $(X,Y) \sim P$. The problem of statistical learning is to construct, for each n, an approximation to f^* on the basis of a *training sequence* $\{Z_i\}_{i=1}^n$, where $Z_i = (X_i, Y_i)$ are i.i.d. according to P, such that this approximation gets better and better as the sample size n tends to infinity. This formulation of the learning problem is referred to as *agnostic* (or *model-free*) learning, reflecting the fact that typically only minimal assumptions are made on the causal relation between X and Y and on the capability of the hypotheses in $\mathcal{F}$ to capture this relation. It is general enough to cover such problems as classification, regression and density estimation.

Formally, a *learning algorithm* (or *learner*, for short) is a sequence $\{\widehat{f}_n\}_{n=1}^\infty$ of maps $\widehat{f}_n : \mathcal{Z}^n \times \mathcal{X} \to \mathcal{Y}$, such that $\widehat{f}_n(Z^n,\cdot) \in \mathcal{F}$ for all n and all $Z^n \in \mathcal{Z}^n$. Let $Z = (X,Y) \sim P$ be independent of the training sequence Z^n. The main quantity of interest is the *generalization error* of the learner,

$$\begin{aligned} L(\widehat{f}_n,P) &\stackrel{\triangle}{=} \mathbb{E}\left[\ell(\widehat{f}_n(Z^n,X),Y)\Big|Z^n\right] \\ &\equiv \int_{\mathcal{Z}} \ell(\widehat{f}_n(Z^n,x),y)dP(x,y). \end{aligned}$$

The generalization error is a random variable, as it depends on the training sequence Z^n. One is chiefly interested in the asymptotic probabilistic behavior of the *excess loss* $L(\widehat{f}_n,P) - L^*(\mathcal{F},P)$ as $n \to \infty$. (Clearly, $L(\widehat{f}_n,P) \geq L^*(\mathcal{F},P)$ for every n.) Under suitable conditions on the loss function ℓ, the hypothesis class $\mathcal{F}$, and the underlying family $\mathcal{P}$ of probability distributions, one can show that there exist learning algorithms which not only *generalize*, i.e., $\mathbb{E}\,L(\widehat{f}_n,P) \to L^*(\mathcal{F},P)$ as $n \to \infty$ for every $P \in \mathcal{P}$ (which is the least one could ask for), but are also *probably approximately correct* (PAC), i.e.

$$\lim_{n\to\infty} P\left(Z^n : L(\widehat{f}_n,P) > L^*(\mathcal{F},P) + \epsilon\right) = 0 \qquad (1)$$

for every $\epsilon > 0$ and every $P \in \mathcal{P}$. (See, e.g., Vidyasagar [2].)

This formulation assumes that the training data are available to the learner with arbitrary precision. This assumption may not always hold, however. For example, the location at which the training data are gathered may be geographically separated from the location where the learning actually takes place. Therefore, the training data may have to be communicated to the learner over a channel of finite capacity. In that case, the learner will see only a quantized version of the training data, and must be able to cope with this to the extent allowed by the fundamental limitations imposed by rate-distortion theory. In this paper, we consider a special case of such learning under rate constraints, when the learner has perfect observation of the input part $X^n = (X_1,\ldots,X_n)$ of the training sequence, while the output part $Y^n = (Y_1,\ldots,Y_n)$ has to be communicated via a noiseless digital channel whose capacity is R bits per sample. This situation, shown in Figure 1, may arise, for example, in remote sensing, where the X_i's are the locations of the sensors and the Y_i's are the measurements of the sensors

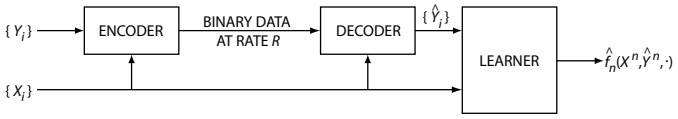

Fig. 1. The set-up for learning from compressed data with side information.

having the form $f_0(X_i) + Z_i$, where $f_0 : \mathcal{X} \to [0,1]$ is some unknown function and the Z_i's are i.i.d. zero-mean Gaussian random variables with variance σ^2. Assuming that the sensors are dispersed at random over some bounded spatial region $\mathcal{X}$ and the location of each sensor is known following its deployment, the task of the sensor array is to deliver, over a rate-limited channel, an approximation $\widehat{Y}^n$ of the measurement vector $Y^n = (Y_1, \ldots, Y_n)$ to some central location, where the vector X^n of the sensor locations and the compressed version $\widehat{Y}^n$ of the sensor measurements will be fed into a learner that will approximate f_0 by some function $\widehat{f}_n(X^n, \widehat{Y}^n, \cdot)$ from a given hypothesis class $\mathcal{F}$.

In this paper, we establish information-theoretic upper bounds on the achievable generalization error in this setting. In particular, we relate the problem of agnostic learning under (partial) rate constraints to conditional rate-distortion theory [3, Section 6.1], [4], [5, Appendix A], which is concerned with lossy source coding in the presence of side information both at the encoder and at the decoder. In the set-up shown in Figure 1, the input part $X^n = (X_1, \ldots, X_n)$ of the training sequence, which is available both to the encoder and to the decoder (hence to the learner), plays the role of the side information, while the output part $Y^n = (Y_1, \ldots, Y_n)$ is to be coded using a lossy source code operating at the rate of R bits per symbol. Furthermore, because the distribution of (X, Y) is known only to be a member of some family $\mathcal{P}$, the lossy codes must be robust in the presence of this uncertainty.

Let us formally state the problem. Let $\mathcal{P}, \mathcal{F}, \ell$ be given. A scheme for agnostic learning under partial rate constraints (from now on, simply a *scheme*) operating at rate R is specified by a sequence of triples $\{(e_n, d_n, \widehat{f}_n)\}_{n=1}^{\infty}$, where $e_n : \mathcal{X}^n \times \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$ is the encoder, $d_n : \mathcal{X}^n \times \{1, \ldots, 2^{nR}\} \to \mathcal{Y}^n$ is the decoder, and $\widehat{f}_n : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{F}$ is the learner. We shall often abuse notation and let $\widehat{f}_n$ denote also the function $\widehat{f}_n(X^n, \widehat{Y}^n, \cdot)$. For each n, the output of the learner is a hypothesis $\widehat{f}_n(X^n, \widehat{Y}^n, \cdot) \in \mathcal{F}$, where $\widehat{Y}^n = d_n(X^n, e_n(X^n, Y^n))$ is the reproduction of Y^n given the side information X^n. For any $P \in \mathcal{P}$, the main object of interest associated with the scheme is the generalization error

$$L(\widehat{f}_n, P) \triangleq \mathbb{E}\left[\ell\big(\widehat{f}_n(X^n, \widehat{Y}^n, X), Y\big)\Big| X^n, Y^n\right],$$

where $(X, Y) \sim P$ is assumed independent of $\{(X_i, Y_i)\}_{i=1}^n$ (to keep the notation simple, we suppress the dependence of the generalization error on the encoder and the decoder). In particular, we are interested in the achievable values of the asymptotic expected excess risk. We say that a pair (R, Δ) is *achievable for* $(\mathcal{F}, \mathcal{P}, \ell)$ if there exists a scheme $\{(e_n, d_n, \widehat{f}_n)\}_{n=1}^{\infty}$ operating at rate R, such that

$$\limsup_{n\to\infty} \mathbb{E}\, L(\widehat{f}_n, P) \le L^*(\mathcal{F}, P) + \Delta$$

for every $P \in \mathcal{P}$. After listing the basic assumptions in Sec. II, we derive in Sec. III sufficient conditions for (R, Δ) to be achievable. We then apply our results to the setting of nonparametric regression in Sec. IV. Discussion of results and an outline of future directions are given in Sec. V.

A. Related work

Previously, the problem of statistical estimation from compressed data was considered by Zhang and Berger [6], Ahlswede and Burnashev [7] and Han and Amari [8] from the viewpoint of multiterminal information theory. In these papers, the underlying family of distributions of (X, Y) is parametric, i.e., of the form $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$, where Θ is a subset of $\mathbb{R}^k$ for some finite k, and one wishes to estimate the "true" parameter θ^*. The i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$ are drawn from P_{θ^*}, and the input part X^n is communicated to the statistician at some rate R_1, while the output part Y^n is communicated at some rate R_2. The present work generalizes to the nonparametric setting the case considered by Ahlswede and Burnashev [7], namely when $R_1 = \infty$. To the best of the author's knowledge, this paper is the first to consider the problem of nonparametric learning from compressed observations with side information.

II. Assumptions

We begin by stating some basic assumptions on $\mathcal{F}$, $\mathcal{P}$ and ℓ. Additional assumptions will be listed in the sequel as needed.

The input space $\mathcal{X}$ is taken to be a measurable subset of $\mathbb{R}^d$, while the output space is either a finite set (as in classification) or the set of reals $\mathbb{R}$ (as in regression or function estimation). We assume throughout that the family $\mathcal{P}$ of distributions on $\mathcal{X} \times \mathcal{Y}$ is such that the mutual information $I(X;Y) < \infty$ for every $P \in \mathcal{P}$. All information-theoretic quantities will be in bits, unless specified otherwise.

We assume that there exists a learning algorithm which generalizes optimally in the absence of any rate constraints. Therefore, our standing assumption on $(\mathcal{F}, \mathcal{P}, \ell)$ will be that the induced function class $\mathcal{L}_\mathcal{F} = \{\ell_f : f \in \mathcal{F}\}$, where $\ell_f(z) \triangleq \ell(f(x), y)$ for all $z = (x, y) \in \mathcal{Z}$, satisfies the *uniform law of large numbers* (ULLN) for every $P \in \mathcal{P}$, i.e.,

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \ell_f(Z_i) - \mathbb{E}\,\ell_f(Z) \right| \to 0, \qquad \text{a.s.} \tag{2}$$

where $Z, Z_1, Z_2, \ldots$ are i.i.d. according to P. Eq. (2) implies that, for any sequence $\{f_n\} \subset \mathcal{F}$,

$$\left| \frac{1}{n}\sum_{i=1}^n \ell_{f_n}(Z_i) - \mathbb{E}\,\ell_{f_n}(Z) \right| \to 0, \qquad \text{a.s.}$$

This holds even in the case when each f_n is random, i.e., $f_n(\cdot) = f_n(Z^n, \cdot)$. The ULLN is a standard ingredient in proofs of consistency of learning algorithms: if $(\mathcal{F}, \mathcal{P}, \ell)$ are

such that (2) holds, then the *Empirical Risk Minimization* algorithm (ERM), given by

$$\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell_f(Z_i),$$

is PAC in the sense of (1) [2, Theorem 3.2].

Next, we assume that the loss function ℓ has the following "generalized Lipschitz" property: there exists a concave, continuous function $\eta : \mathbb{R}^+ \to \mathbb{R}^+$, such that for all $f \in \mathcal{F}$, $x \in \mathcal{X}$ and $u, u' \in \mathcal{Y}$

$$|\ell(f(x), u) - \ell(f(x), u')| \le \eta(\ell(u, u')). \tag{3}$$

This holds, for example, in the following cases:

- Suppose that ℓ is a metric on $\mathcal{Y}$. Then, by the triangle inequality we have $\ell(y, u) \le \ell(y, u') + \ell(u', u)$ for all $y, u, u' \in \mathcal{Y}$, so (3) holds with $\eta(t) = t$.
- Suppose that $\mathcal{Y} = [0, 1]$ and $\ell(u, u') = |u - u'|^p$ for some $p \ge 1$. Then one can show that

  $$|\ell(f(x), u) - \ell(f(x), u')| \le p|u - u'|$$

  for all $f : \mathcal{X} \to \mathcal{Y}$, $x \in \mathcal{X}$ and $u, u' \in \mathcal{Y}$, so (3) holds with $\eta(t) = pt^{1/p}$.

Finally, we need to pose some assumptions on the metric structure of the class $\mathcal{P}$ with respect to the *variational distance* [9, Sec. 5.2], which for any two probability distributions P_1, P_2 on a measurable space $(\mathcal{Z}, \mathcal{A})$ is defined by

$$d_V(P_1, P_2) \triangleq 2 \sup_{A\in\mathcal{A}} |P_1(A) - P_2(A)|.$$

A finite set $\{P_1, \ldots, P_M\} \subset \mathcal{P}$ is called an *(internal) ϵ-net* for $\mathcal{P}$ with respect to d_V if

$$\sup_{P\in\mathcal{P}} \min_{1\le m\le M} d_V(P, P_m) \le \epsilon.$$

The cardinality of a minimal ϵ-net, denoted by $N(\epsilon, \mathcal{P})$, is called the *ϵ-covering number* of $\mathcal{P}$ w.r.t. d_V, and the *Kolmogorov ϵ-entropy* of $\mathcal{P}$ is defined as $H(\epsilon, \mathcal{P}) \triangleq \log N(\epsilon, \mathcal{P})$ [10]. We assume that the class $\mathcal{P}$ satisfies *Dobrushin's entropy condition* [11], i.e., for every $c > 0$

$$\lim_{\epsilon\to 0} \frac{H(\epsilon, \mathcal{P})}{2^{c/\epsilon}} = 0. \tag{4}$$

This condition is satisfied, for example, in the following cases: (1) $\mathcal{X}$ and $\mathcal{Y}$ are both finite sets; (2) $\mathcal{P}$ is a finite family; (3) $\mathcal{Z}$ is a compact subset of a Euclidean space, and all $P \in \mathcal{P}$ are absolutely continuous with densities satisfying a uniform Lipschitz condition [10], [11].

III. The results

To state our results we shall need some notions from conditional rate-distortion theory [3, Sec. 6.1], [4], [5, Appendix A]. Fix some $P \in \mathcal{P}$. Given a pair $(X, Y) \sim P$ and a nonnegative real number D, define the set $\mathcal{M}(D)$ to consist of all $\mathcal{Y}$-valued random variables $\widehat{Y}$ jointly distributed with (X, Y) and satisfying the constraint $\mathbb{E}\,\ell(Y, \widehat{Y}) \le D$, where the expectation is taken with respect to the joint distribution of $X, Y, \widehat{Y}$. Then the *conditional rate-distortion function* of Y given X w.r.t. P is defined by

$$R_{Y|X}(D, P) \triangleq \inf\left\{ I(Y; \widehat{Y}|X) : \widehat{Y} \in \mathcal{M}(D) \right\},$$

where $I(Y; \widehat{Y}|X)$ is the conditional mutual information between Y and $\widehat{Y}$ given X. Our assumption that $I(X; Y) < \infty$ ensures the existence of $R_{Y|X}(D, P)$ [5]. In operational terms, $R_{Y|X}(D, P)$ is the minimum number of bits needed to describe Y with expected distortion of at most D given perfect observation of a correlated random variable X (the side information) when $(X, Y) \sim P$. As a function of D, $R_{Y|X}(D, P)$ is convex and strictly decreasing everywhere it is finite, hence it is invertible. The inverse function is called the *conditional distortion-rate function* of Y given X and is denoted by $D_{Y|X}(R, P)$. Finally, let

$$\mathbb{D}_{Y|X}(R, \mathcal{P}) \triangleq \sup_{P\in\mathcal{P}} D_{Y|X}(R, P).$$

We assume that $\mathbb{D}_{Y|X}(R, \mathcal{P}) < \infty$ for all $R \ge 0$.

We shall also need the following lemma, which can be proved by a straightforward extension of Dobrushin's random coding argument from [11] to the case of side information available to the encoder and to the decoder:

Lemma 3.1. Let $\mathcal{P}$ satisfy Dobrushin's entropy condition (4). Assume that the loss function ℓ either is bounded or satisfies a uniform moment condition

$$\sup_{P\in\mathcal{P}} \mathbb{E}[\ell(Y, y_0)^{1+\delta}] < \infty \tag{5}$$

for some $\delta > 0$ with respect to some fixed reference letter $y_0 \in \mathcal{Y}$. Then for every rate $R \ge 0$ there exists a sequence $\{(e_n, d_n)\}_{n=1}^{\infty}$ of encoders $e_n : \mathcal{X}^n \times \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$ and decoders $d_n : \mathcal{X}^n \times \{1, \ldots, 2^{nR}\} \to \mathcal{Y}^n$, such that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{E}\,\ell_n(Y^n, \widehat{Y}^n) \le \mathbb{D}_{Y|X}(R, \mathcal{P}),$$

where $\widehat{Y}^n = d_n(X^n, e_n(X^n, Y^n))$ and $\ell_n(Y^n, \widehat{Y}^n) = n^{-1}\sum_{i=1}^{n} \ell(Y_i, \widehat{Y}_i)$ is the normalized cumulative loss between Y^n and $\widehat{Y}^n$.

Our main result can then be stated as follows:

Theorem 3.1. Under the stated assumptions, for any $R \ge 0$ there exists a scheme $\{(e_n, d_n, \widehat{f}_n)\}$ operating at rate R, such that

$$\limsup_{n\to\infty} \mathbb{E}\, L(\widehat{f}_n, P) \le L^*(\mathcal{F}, P) + 2\eta(\mathbb{D}_{Y|X}(R, \mathcal{P})).$$

Thus, $(R, 2\eta(\mathbb{D}_{Y|X}(R, \mathcal{P})))$ is achievable for every $R \ge 0$.

Proof: Given n, $Z^n \in \mathcal{Z}^n$ and $f \in \mathcal{F}$, define the *empirical risk*

$$\widehat{L}_{Z^n}(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell_f(Z^n)$$

and the minimum empirical risk

$$\widehat{L}^*_{Z^n}(\mathcal{F}) \triangleq \inf_{f\in\mathcal{F}} \widehat{L}_{Z^n}(f).$$

We shall write $\widehat{L}_{X^n,Y^n}(f)$ and $\widehat{L}^*_{X^n,Y^n}(\mathcal{F})$ whenever we need to emphasize separately the roles of X^n and Y^n.

Suppose that the encoder e_n and the decoder d_n are given. Let $\widehat{Y}^n$ denote the reproduction of Y^n given the side information X^n, i.e., $\widehat{Y}^n = d_n(X^n, e_n(X^n, Y^n))$. We then define our learner $\widehat{f}_n$ by

$$\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \widehat{L}_{X^n,\widehat{Y}^n}(f). \tag{6}$$

In other words, having received the side information X^n and the reproduction $\widehat{Y}^n$, the learner performs ERM over $\mathcal{F}$ on $\{(X_i, \widehat{Y}_i)\}_{i=1}^n$. Using the property (3) of the loss function ℓ and the concavity of η, we have the following estimate:

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left|\widehat{L}_{X^n,Y^n}(f) - \widehat{L}_{X^n,\widehat{Y}^n}(f)\right| \\
&\quad\le \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \left|\ell(f(X_i), Y_i) - \ell(f(X_i), \widehat{Y}_i)\right| \\
&\quad\le \frac{1}{n}\sum_{i=1}^n \eta(\ell(Y_i, \widehat{Y}_i)) \\
&\quad\le \eta\big(\ell_n(Y^n, \widehat{Y}^n)\big).
\end{aligned} \tag{7}$$

In particular, this implies that

$$\left|\widehat{L}_{X^n,Y^n}(\widehat{f}_n) - \widehat{L}_{X^n,\widehat{Y}^n}(\widehat{f}_n)\right| \le \eta\big(\ell_n(Y^n, \widehat{Y}^n)\big) \tag{8}$$

and

$$\left|\widehat{L}^*_{X^n,Y^n}(\mathcal{F}) - \widehat{L}^*_{X^n,\widehat{Y}^n}(\mathcal{F})\right| \le \eta\big(\ell_n(Y^n, \widehat{Y}^n)\big). \tag{9}$$

We then have

$$\begin{aligned}
\widehat{L}_{X^n,Y^n}(\widehat{f}_n) &\overset{(a)}{\le} \widehat{L}_{X^n,\widehat{Y}^n}(\widehat{f}_n) + \eta\big(\ell_n(Y^n, \widehat{Y}^n)\big) \\
&\overset{(b)}{=} \widehat{L}^*_{X^n,\widehat{Y}^n}(\mathcal{F}) + \eta\big(\ell_n(Y^n, \widehat{Y}^n)\big) \\
&\overset{(c)}{\le} \widehat{L}^*_{X^n,Y^n}(\mathcal{F}) + 2\eta\big(\ell_n(Y^n, \widehat{Y}^n)\big),
\end{aligned}$$

where (a) follows from (8), (b) from the definition of $\widehat{f}_n$, and (c) from (9). Suppose that the data are distributed according to a particular $P \in \mathcal{P}$. Taking expectations and using the concavity of η and Jensen's inequality, we obtain

$$\mathbb{E}\,\widehat{L}_{Z^n}(\widehat{f}_n) \le \mathbb{E}\,\widehat{L}^*_{Z^n}(\mathcal{F}) + 2\eta\big(\mathbb{E}\,\ell_n(Y^n, \widehat{Y}^n)\big).$$

Using this bound and the continuity of η, we can write

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{E}\,L(\widehat{f}_n, P) - L^*(\mathcal{F}, P) \\
&\quad\le \lim_{n\to\infty} \mathbb{E}\left[L(\widehat{f}_n, P) - \widehat{L}_{Z^n}(\widehat{f}_n)\right] \\
&\qquad + \lim_{n\to\infty} \mathbb{E}\left[\widehat{L}^*_{Z^n}(\mathcal{F}) - L^*(\mathcal{F}, P)\right] \\
&\qquad + 2\eta\Big(\limsup_{n\to\infty} \mathbb{E}\,\ell_n(Y^n, \widehat{Y}^n)\Big).
\end{aligned} \tag{10}$$

The two leading terms on the right-hand side of this inequality are zero by the ULLN. Moreover, given R, Lemma 3.1 asserts the existence of a sequence $\{(e_n, d_n)\}_{n=1}^\infty$ of encoders $e_n : \mathcal{X}^n \times \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$ and decoders $d_n : \mathcal{X}^n \times \{1, \ldots, 2^{nR}\} \to \mathcal{Y}^n$, such that

$$\limsup_{n\to\infty} \mathbb{E}\,\ell_n(Y^n, \widehat{Y}^n) \le \mathbb{D}_{Y|X}(R, \mathcal{P}), \qquad \forall P \in \mathcal{P}.$$

Substitution of this into (10) proves the theorem. ∎

Corollary 3.2. All pairs (R, Δ) with $\Delta \ge 2\eta(\mathbb{D}_{Y|X}(R, \mathcal{P}))$ are achievable.

Remark 3.1. In the Appendix, we show that a corresponding lower bound derived by the usual methods for proving converses in lossy source coding is strictly weaker than the "obvious" lower bound based on the observation that $\mathbb{E}\,L(\widehat{f}_n, P) \ge L^*(\mathcal{F}, P)$ for any $\widehat{f}_n$. It may be possible to obtain nontrivial lower bounds in the minimax setting, which we leave for future work (see also Sec. V).

Remark 3.2. Under some technical conditions on the function class $\{\ell_f : f \in \mathcal{F}\}$ (see, e.g., [12]), one can show that

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left|\widehat{L}_{Z^n}(f) - L(f, P)\right| \le C/\sqrt{n}, \qquad \forall P \in \mathcal{P}$$

for some constant C that depends on $\mathcal{F}, \ell$. Using this fact and the same bounding method that led to Eq. (10), but without taking the limit superior, we can get the following finite-sample bound for every scheme $\{(e_n, d_n, \widehat{f}_n)\}_{n=1}^\infty$ with $\widehat{f}_n$ given by (6) and *arbitrary* e_n, d_n:

$$\mathbb{E}\,L(\widehat{f}_n, P) \le L^*(\mathcal{F}, P) + 2\eta\big(\mathbb{E}\,\ell_n(Y^n, \widehat{Y}^n)\big) + C'/\sqrt{n},$$

where $C' = 2C$.

The following theorem shows that we can replace condition (3) with the requirement that ℓ be a power of a metric:

Theorem 3.3. Suppose that the loss function ℓ is of the form $\ell(y, u) = d(y, u)^r$ for some $r \ge 1$, where d is a metric on $\mathcal{Y}$. Then for any rate $R \ge 0$ the scheme constructed in the proof of Theorem 3.1 is such that

$$\limsup_{n\to\infty} \mathbb{E}\left[L(\widehat{f}_n, P)^{1/r}\right] \le L^*(\mathcal{F}, P)^{1/r} + 2\mathbb{D}_{Y|X}(R, \mathcal{P})^{1/r}$$

holds for every $P \in \mathcal{P}$.

Proof: We proceed essentially along the same lines as in the proof of Theorem 3.1, except that the bound (7) is replaced with an argument based on Minkowski's inequality to yield

$$\mathbb{E}\left[\widehat{L}_{Z^n}(\widehat{f}_n)^{1/r}\right] \le \mathbb{E}\left[\widehat{L}^*_{Z^n}(\mathcal{F})^{1/r}\right] + 2\Big(\mathbb{E}\,\ell_n(Y^n, \widehat{Y}^n)\Big)^{1/r}.$$

The rest is immediate using the ULLN as well as concavity and continuity of $t \mapsto t^{1/r}$ for $t \ge 0$. ∎

IV. Example: nonparametric regression

As an example, let us consider the setting of nonparametric regression. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. The training data are of the form

$$Y_i = f_0(X_i) + Z_i, \qquad 1 \le i \le n \tag{11}$$

where the regression function f_0 belongs to some specified class $\mathcal{F}$ of functions from $\mathcal{X}$ into $[0,1]$, the X_i's are i.i.d. random variables drawn from the uniform distribution on $\mathcal{X}$, and the Z_i's are i.i.d. zero-mean normal random variables with variance σ^2, independent of X^n. We take $\ell(y,u) = |y-u|^2$, the squared loss. Note that ℓ satisfies the condition of Theorem 3.3 with $r=2$.

Because f_0 is unknown, we take as the underlying family $\mathcal{P}$ the class of all absolutely continuous distributions with densities of the form $p_f(x,y) = V^{-1}\mathcal{N}(y; f(x), \sigma^2)$, $f \in \mathcal{F}$, where V is the volume of $\mathcal{X}$ and $\mathcal{N}(y; f(x), \sigma^2)$ is the one-dimensional normal density with mean $f(x)$ and variance σ^2. Because the functions in $\mathcal{F}$ are bounded between 0 and 1, it is easy to show that the uniform moment condition (5) of Lemma 3.1 is satisfied with $\delta = 1$ and $y_0 = 0$.

We suppose that ℓ and $\mathcal{F}$ are such that the function class $\mathcal{L}_{\mathcal{F}}$ satisfies the ULLN.[1] Let Q denote the uniform distribution on $\mathcal{X}$ and for any square-integrable function f on $\mathcal{X}$ define the L_2 norm by

$$\|f\|_{2,Q}^2 \stackrel{\triangle}{=} \int_{\mathcal{X}} f^2(x) dQ(x) \equiv \frac{1}{V}\int_{\mathcal{X}} f^2(x)dx.$$

Let us denote by $N_{2,Q}(\epsilon, \mathcal{F})$ the ϵ-covering number of $\mathcal{F}$ w.r.t. $\|\cdot\|_{2,Q}$, i.e., the smallest number M such that there exist M functions $\{f_m\}_{m=1}^M$ in $\mathcal{F}$ satisfying

$$\sup_{f\in\mathcal{F}} \min_{1\le m\le M} \|f - f_m\|_{2,Q} \le \epsilon.$$

We assume that $\mathcal{F}$ is such that for every $c > 0$

$$\lim_{\epsilon\to 0} \frac{\log N_{2,Q}(\epsilon, \mathcal{F})}{2^{c/\epsilon}} = 0. \tag{12}$$

This condition holds, for example, if the functions in $\mathcal{F}$ are uniformly Lipschitz or if $\mathcal{X}$ is a bounded interval in $\mathbb{R}$ and $\mathcal{F}$ consists of functions satisfying a Sobolev-type condition [10].

Lemma 4.1. If $\mathcal{F}$ satisfies (12), then $\mathcal{P}$ satisfies Dobrushin's entropy condition (4).

Proof: Given $f \in \mathcal{F}$, let P_f denote the distribution with the density p_f. It is straightforward to show that

$$I(P_f\|P_g) = \frac{1}{2\sigma^2}\|f-g\|_{2,Q}^2, \qquad \forall f,g \in \mathcal{F}$$

where $I(\cdot\|\cdot)$ is the relative entropy (information divergence) between two probability distributions, in nats. Using Pinsker's inequality $d_V(P_1,P_2) \le \sqrt{2I(P_1\|P_2)}$ [9, Lemma 5.2.8], we get

$$d_V(P_f\|P_g) \le \frac{1}{\sigma}\|f-g\|_{2,Q}, \qquad \forall f,g\in\mathcal{F}. \tag{13}$$

Given $\epsilon > 0$, let $\{f_m\}_{m=1}^M \subset \mathcal{F}$ be a $\sigma\epsilon$-net for $\mathcal{F}$ w.r.t. $\|\cdot\|_{2,Q}$. Then from (13) it follows that

$$\sup_{f\in\mathcal{F}} \min_{1\le m\le M} d_V(P_f, P_{f_m}) \le \sup_{f\in\mathcal{F}} \min_{1\le m\le M} \frac{\|f-f_m\|_{2,Q}}{\sigma} \le \epsilon,$$

i.e., $\{P_{f_m}\}_{m=1}^M$ is an ϵ-net for $\mathcal{P}$ w.r.t. d_V. This implies, in particular, that $N(\epsilon,\mathcal{P}) \le N_{2,Q}(\sigma\epsilon,\mathcal{F})$ for every $\epsilon > 0$. This, together with (12), proves the lemma. ∎

Lemma 4.2. For any $R \ge 0$, $\mathbb{D}_{Y|X}(R,\mathcal{P}) = \sigma^2 2^{-2R}$.

Proof: Fix some $f \in \mathcal{F}$ and consider a pair $(X,Y) \sim P_f$. Then $Y = f(X) + Z$, where $Z \sim \mathrm{Normal}(0,\sigma^2)$ is independent of X. Because ℓ is a difference distortion measure, Theorem 7 of [4] says that, for any measurable function $\psi : \mathcal{X} \to \mathcal{Y}$,

$$D_{Y|X}(R,P_f) = D_{Y-\psi(X)|X}(R, P_{f-\psi}),$$

where $P_{f-\psi}$ is the distribution of

$$Y - \psi(X) \equiv f(X) - \psi(X) + Z;$$

furthermore, if $Y - \psi(X)$ is independent of X, then $D_{Y|X}(R,P_f) = D_{Y-\psi(X)}(R)$, the (unconditional) distortion-rate function of $Y - \psi(X)$. Taking $\psi = f$, we get $D_{Y|X}(R,P_f) = D(R,\sigma^2)$, the distortion-rate function of a memoryless Gaussian source with variance σ^2 w.r.t. squared error loss, which is equal to $\sigma^2 2^{-2R}$ [3, Theorem 9.3.2]. Hence $D_{Y|X}(R,P_f)$ is independent of f. Taking the supremum over $\mathcal{F}$ finishes the proof. ∎

Now we can state and prove the main result of this section:

Theorem 4.1. Consider the regression setting of (11). Under the stated assumptions, for any $R \ge 0$ there exists a scheme $\{(e_n, d_n, \widehat{f}_n)\}_{n=1}^\infty$, such that

$$\limsup_{n\to\infty} \mathbb{E}\left[L(\widehat{f}_n, P_f)^{1/2}\right] \le \sigma(1 + 2^{-R+1}) \tag{14}$$

holds for every $f \in \mathcal{F}$.

Proof: As follows from the above, the triple $(\mathcal{F},\mathcal{P},\ell)$ satisfies all the assumptions of Theorem 3.3. Therefore for any $R \ge 0$ there exists a scheme $\{(e_n,d_n,\widehat{f}_n)\}_{n=1}^\infty$ operating at rate R, such that

$$\limsup_{n\to\infty} \mathbb{E}\left[L(\widehat{f}_n, P_f)^{1/2}\right] \le L^*(\mathcal{F},P_f)^{1/2} + 2^{-R+1}\sigma, \tag{15}$$

holds for every $f \in \mathcal{F}$ (we have also used Lemma 4.2). It is not hard to show that

$$L(g,P_f) = \|f-g\|_{2,Q}^2 + \sigma^2, \qquad \forall f,g\in\mathcal{F},$$

whence it follows that $L^*(\mathcal{F},P_f) = \sigma^2$ for every $f \in \mathcal{F}$. Substituting this into (15), we get (14). ∎

V. Discussion and future work

We have derived information-theoretic bounds on the achievable generalization error in learning from compressed data (with side information). There is a close relationship between this problem and the theory of robust lossy source coding with side information at the encoder and the decoder. A major difference between this setting and the usual setting of learning theory is that the techniques are no longer *distribution-free* because restrictions must be placed on the underlying family of distributions in order to guarantee the

[1] See Györfi et al. [13] for a detailed exposition of the various conditions when this is true.

existence of a suitable source code. The theory was applied to the problem of nonparametric regression in Gaussian noise, where we have shown that the penalty incurred for using compressed observations decays exponentially with the rate.

We have proved Theorems 3.1 and 3.3 by adopting ERM as our learning algorithm and optimizing the source code to deliver the best possible reconstruction of the training data. In effect, this imposes a *separation structure* between learning and source coding. While this "modular" approach is simplistic (clearly, additional performance gains could be attained by designing the encoder, the decoder and the learner jointly), it may be justified in such applications as remote sensing. For instance, if the source code and the learner were designed jointly, then any change made to the hypothesis class (say, if we decided to replace the currently used hypothesis class with another based on tracking the prior performance of the network) might call for a complete redesign of the source code and the sensor network, which may be a costly step. With the modular approach, no such redesign is necessary: one merely makes the necessary adjustments in the learning algorithm, while the sensor network continues to operate as before.

Let us close by sketching some directions for future work. First of all, it would be of interest to derive information-theoretic lower bounds on the generalization performance of rate-constrained learning algorithms. In particular, just as Ahlswede and Burnashev had done in the parametric case [7], we could study the asymptotics of the *ninimax excess risk*

$$\delta_n(R) \stackrel{\triangle}{=} \inf_{(e_n,d_n,\widehat{f}_n)} \sup_{P\in\mathcal{P}} \Big[\mathbb{E}\, L(\widehat{f}_n, P) - L^*(\mathcal{F}, P) \Big],$$

where the infimum is over all encoders, decoders and learners operating on a length-n training sequence at rate R. Secondly, we could dispense with the assumption that the learner has perfect observation of the input part of the training sample, in analogy to the situation dealt with by Zhang and Berger [6]. Finally, keeping in mind the motivating example of sensor networks, it would be useful to replace the block coding approach used here with an efficient distributed scheme.

ACKNOWLEDGMENTS

Discussions with Todd Coleman are gratefully acknowledged. This work was supported by the Beckman Fellowship.

APPENDIX

Let us assume for simplicity that $\mathcal{P}$ is a singleton, $\mathcal{P} = \{P\}$, and that $\mathcal{Y}$ is a finite set. Consider a scheme $\{(e_n, d_n, \widehat{f}_n)\}$ operating at rate R. Fix n and define the n-tuple W^n via

$$W_i \stackrel{\triangle}{=} \widehat{f}_n(X^n, \widehat{Y}^n, X_i), \qquad 1 \le i \le n.$$

Also, let $J = e_n(X^n, Y^n)$. Then we can write

$$\begin{aligned}
nR &\ge H(J|X^n) \\
&\ge H(\widehat{Y}^n|X^n) \\
&\ge I(\widehat{Y}^n; Y^n|X^n) \\
&= H(Y^n|X^n) - H(Y^n|X^n, \widehat{Y}^n) \\
&= H(Y^n|X^n) - H(Y^n|X^n, \widehat{Y}^n, W^n) \quad \text{(A.1)} \\
&= \sum_{i=1}^n [H(Y_i|X_i) - H(Y_i|X^n, \widehat{Y}^n, W^n, Y^{i-1})] \\
&\ge \sum_{i=1}^n [H(Y_i|X_i) - H(Y_i|X_i, W_i)] \\
&= \sum_{i=1}^n I(Y_i; W_i|X_i) \\
&\ge \sum_{i=1}^n R_{Y|X}(\mathbb{E}\,\ell(W_i, Y_i), P) \\
&\ge nR_{Y|X}(\mathbb{E}\,\ell_n(W^n, Y^n), P),
\end{aligned}$$

where (A.1) follows from the fact that W^n is a function of $\widehat{Y}^n$ and X^n. The remaining steps follow from standard information-theoretic identities and from convexity. Therefore,

$$\liminf_{n\to\infty} \mathbb{E}\,\ell_n(W^n, Y^n) \ge D_{Y|X}(R, P).$$

Because $\mathbb{E}\, L(\widehat{f}_n, P) = \mathbb{E}\,\ell_n(W^n, Y^n) + o(1)$ by the ULLN,

$$\liminf_{n\to\infty} \mathbb{E}\, L(\widehat{f}_n, P) \ge D_{Y|X}(R, P). \quad \text{(A.2)}$$

Now, given any $f \in \mathcal{F}$, we can interpret $f(X)$ as a zero-rate approximation of Y (using only the side information X), so $L(f, P) \ge D_{Y|X}(0, P) \ge D_{Y|X}(R, P)$ for any $R \ge 0$. In particular, $L^*(\mathcal{F}, P) \ge D_{Y|X}(R, P)$ for all R, and

$$\liminf_{n\to\infty} \mathbb{E}\, L(\widehat{f}_n, P) \ge L^*(\mathcal{F}, P) \ge D_{Y|X}(R, P)$$

for all R. Thus, the information-theoretic lower bound (A.2) is weaker than the bound $\liminf_{n\to\infty} \mathbb{E}\, L(\widehat{f}_n, P) \ge L^*(\mathcal{F}, P)$.

REFERENCES

[1] D. Haussler, "Decision-theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Comput.*, vol. 100, pp. 78–150, 1992.

[2] M. Vidyasagar, *Learning and Generalization*, 2nd ed. London: Springer-Verlag, 2003.

[3] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[4] R. M. Gray, "Conditional rate-distortion theory," Stanford Electronics Laboratories, Tech. Rep. 6502-2, 1972.

[5] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: general sources," *Inform. Control*, vol. 38, pp. 60–80, 1978.

[6] Z. Zhang and T. Berger, "Estimation via compressed information," *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 198–211, March 1988.

[7] R. Ahlswede and M. V. Burnashev, "On minimax estimation in the presence of side information about remote data," *Ann. Statist.*, vol. 18, no. 1, pp. 141–171, 1990.

[8] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2300–2324, October 1998.

[9] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[10] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," in *Amer. Math. Soc. Transl.*, ser. 2, 1961, vol. 17, pp. 277–364.

[11] R. L. Dobrushin, "Unified methods for optimal quantization of messages," in *Problemy Kibernetiki*, A. A. Lyapunov, Ed. Moscow: Nauka, 1970, vol. 22, pp. 107–156, in Russian.

[12] S. Mendelson, "A few notes on statistical learning theory," in *Advanced Lectures in Machine Learning*, ser. Lecture Notes in Computer Science, S. Mendelson and A. J. Smola, Eds. Springer, 2003, vol. 2600, pp. 1–40.

[13] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer, 2002.